



Paper AS05

How GSK Tech is Democratizing Data Science

Pragathi Kotha Venkata, GlaxoSmithKline, Durham, United States

ABSTRACT

As a part of the Digital transformation journey, one of the goals of GSK technology organization is to Democratize Data science. As it wouldn't be possible to have a data scientist in every room to enable that, we needed to scale our skillset to develop citizen data scientists across the company. Our guiding principle is to ask-analyze-act. We started by deploying a user research survey to ask about existing pain points. We analyzed the result and devised a strategy to act on. Our MVP (Minimal Viable Product) objectives were: 1. Provision of modern self-service Data science stack - an easily accessible way for requesting data science tools in a self-service fashion; 2. Inspiring code reuse and collaboration - by promoting usage of platforms such as GitHub and Guilds on workplace; 3. Training and development - leveraging platforms such as Pluralsight, Learnathons & Hackathons.

INTRODUCTION

As a new enterprise data platform group, our first action was to deploy a user research survey using an internal survey tool within our VPN called GSK Select survey. Find below the survey questions which were used.

Data Science User Research Survey: Core Technology

Page 1 of 1

Data Analysts, Scientist & Enthusiasts; the Data Product and Engineering team would love to hear from you to learn more about the tools you use at GSK for Data Science and Analysis. This short survey will take less than 2 minutes, and will be used to ensure we make user centric decisions around Data Science tools and capabilities.

1. Overall, how happy are you with current tools and technologies available to you for carrying out data and analytics tasks?

☐ Very satisfied

☐ Satisfied

☐ Neither satisfied or dissatisfied

☐ Dissatisfied

☐ Very Dissatisfied

2. What is the primary reason for your reply above?

3. On an average, what percentage of your work involves working with data - analyzing data or cleaning data or maintaining data and databases, etc?

☐ 25% or less
☐ More than 25% but less than 50%
☐ Over 50%
☐ My day job is working with data
☐ I don't work with data

4. What is your level of expertise with the following tools/languages?

	Never used it	Beginner	Intermediate	Advanced
SAS	☐	☐	☐	☐
SQL	☐	☐	☐	☐
Python	☐	☐	☐	☐
R	☐	☐	☐	☐
Spotfire	☐	☐	☐	☐
PowerBI	☐	☐	☐	☐
Qlik	☐	☐	☐	☐
Tableau	☐	☐	☐	☐

5. Please list any other tools you use that are not listed above.

6. If you use excel or other similar tools for your data and analytics jobs, would you like to learn other tools/technologies which can provide you with additional capabilities and skills?

☐ Yes
☐ No
☐ Not applicable

7. If you answered yes to the question above, what are other tools and technologies that you are interested in exploring and learning more?

☐ Advanced data modeling tools
☐ Machine learning tools
☐ Advanced Visualization tools
☐ Interested but I'm not sure
☐ Other, please specify

8. Please provide any other suggestions or comments regarding tools, technologies or platforms that will help improve your ability to work with data at GSK.

9. Which Organization do you belong to?

- ☐ Pharma R&D
- ☐ Pharmaceuticals
- ☐ GMS
- ☐ Vaccines
- ☐ Consumer Health
- ☐ Viiv
- ☐ Global Tech Organization
- ☐ Other, please specify

10. How do you identify yourself?

- ☐ Data Analyst
- ☐ Business Analyst
- ☐ Data Management Professional
- ☐ Data Engineer
- ☐ Data Scientist
- ☐ Data Science Enthusiast
- ☐ Other, please specify

11. What is your experience with data and analytics in years?

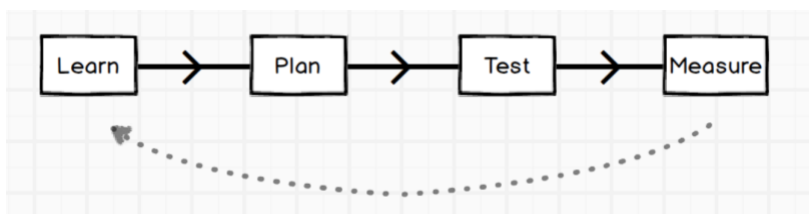
- ☐ I don't work with data or analytics
- ☐ Less than 1 year
- ☐ 1 to 3 years
- ☐ 3 to 10 years
- ☐ More than 10 years

Done

We ran the survey for two weeks and capped at 100 responses. The major themes we came across in our analysis of the survey results were:

1. Lack of access to tools across groups (Some teams have better access, others not much)
2. Need for training and awareness around capabilities

1. PROVISIONING OF MODERN DATA SCIENCE STACK



To address the pain point of lack of access to Data science tooling across the enterprise we needed to:

- a) Define MVP (Minimum Viable Product) Tool Strategy: Identify the tools majority will benefit from
- b) Enable a mechanism for self-service:
- c) Have a streamlined enterprise licensing process (where possible)

MVP TOOL STRATEGY

To define our MVP tool strategy, we first identified our initial customer personas as: Data Analyst & Data Scientists.

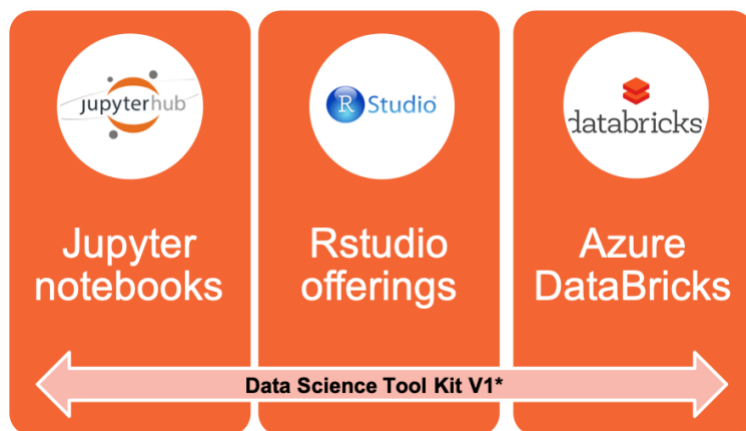
Define Initial Personas



We then identified the common steps in the Data science/ analysis lifecycle which are Preparation, Analysis, Reflection, and Dissemination. Our tools were evaluated to ensure usability and efficiency for these phases.

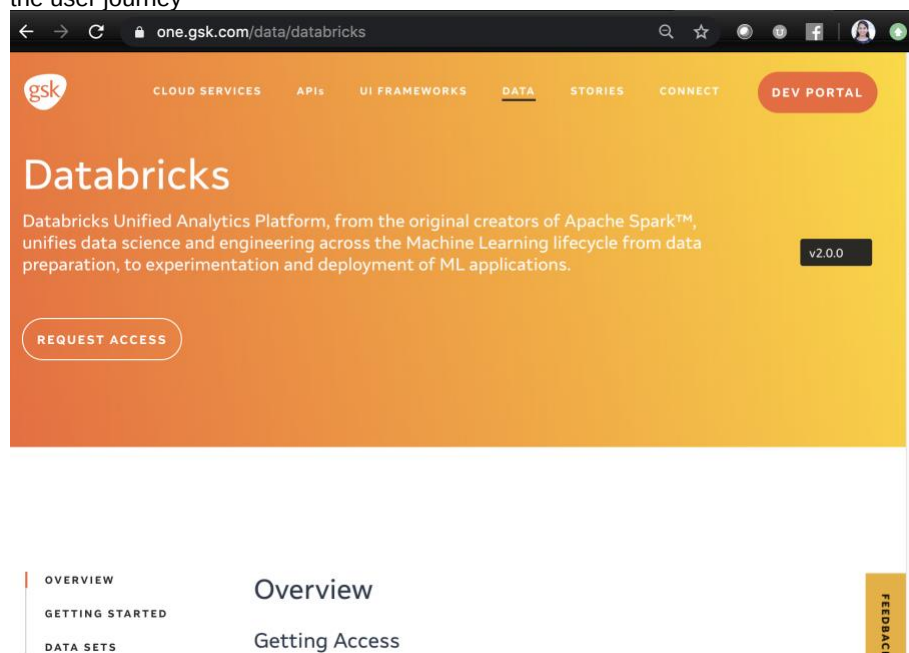
Our final tool choices were complemented by subject matter expertise, existing collaborations and strategic partnerships. Our MVP Data Science tool kit consisted of:

- Azure Databricks® (<https://azure.microsoft.com/en-us/services/databricks/>)
- RStudio® offerings (RStudio server (<https://rstudio.com/products/rstudio/>), RStudio Connect(<https://rstudio.com/products/connect/>)
- Jupyter® notebooks (<https://jupyter.org/>)



SELF-SERVICE

To empower Data scientists and developers to build products at speed, One.gsk.com was launched to serve as a one stop shop where users can request Data science tooling in a self-service fashion. See below the snapshot of the solution describing the user journey



Request to try Databricks in a sandbox environment

Complete this form and we'll be in touch with the credentials you need to get started.

NAME	EMAIL
Please enter your name*	Please enter your email address*
MUD ID	BUSINESS UNIT
Please enter your Mud ID *	Select

Do you have a dataset you'd like to use?

☒ No ☐ Yes

☐ By submitting the request, you agree to the [ONE Platform Data, Product, and Engineering Shared Responsibilities Agreement](#).

[CANCEL](#)

CENTRALIZED LICENSING MODEL

One of the barriers we encountered in enabling democratization of data science tooling was the lack of enterprise licenses of commonly used tools.

ENTERPRISE R LICENSE

Of the ubiquitously used software, RStudio offerings were something we decided to handle first. RStudio had contracts with different groups within our company. Some groups had access to all of its offerings whereas others struggled to get budget to use any of their products. By procuring an enterprise licenses we not only enabled access across the company but were able to get significant savings as well. It was a win-win for even the vendor as they didn't have to have multiple legal contracts with different groups. Our procurement colleagues were extremely helpful in this endeavor.

By having a centralized licensing model where possible, opportunities like these can be realized even more.

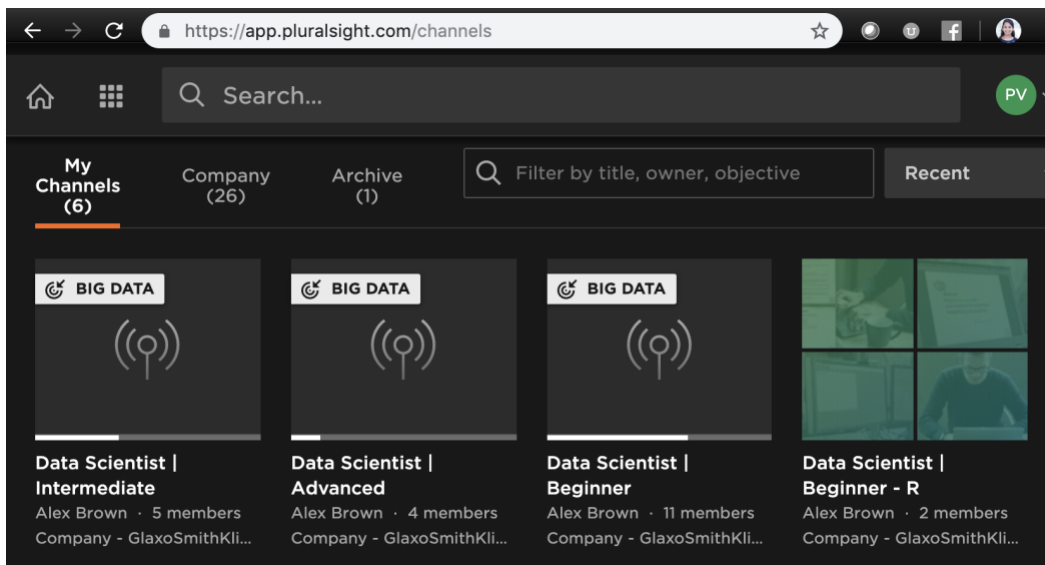
2. TRAINING AND DEVELOPMENT

Another key barrier to realize our ambitions was the need for increased training. Many of those whom we surveyed had basic data skills and very interested to expand their skill set but did not have the necessary tools to do so. The avenues through which we are currently providing training and development are:

- Self-paced learning through Pluralsight
- In person sessions through Learnathons
- Outside in collaboration such as Women in Data Science initiatives

PLURALSIGHT

Pluralsight(<https://app.pluralsight.com/library/>) was selected to pilot as a learning platform. We leveraged subject matter experts across the company to develop pathways and select courses of choice that add maximum value. Licenses for pilots were shared with different groups to collect feedback. Based on extremely positive feedback, we recently chose it as our learning platform for self-paced learning.



HANDS ON LEARNATHONS

Beyond self-paced learning, we got requests for in person trainings from many members. We initially piloted with Databricks training in conjunction with Python training. We had about 100 attendance in the very first month of launching the training. There was interest in expanding across different geographic locations. To address the demand, we launched “Train the trainer” program with the intention of training one or two representatives from each site who could organize future Learnathons at their site. Current sites where Learnathons were conducted are Raleigh area, Philadelphia area and Poland. We have interested “train the trainer” participants across Poland, Malaysia, Belgium, India and UAE.



OUTREACH

Apart from internal sessions we focused on bringing in outside-in perspectives, work with local academia and connect with other peers in local areas to learn more about opportunities and challenges. WiDS Durham is one such event GSK is proud to sponsor. The Women in Data Science (WiDS) initiative(<https://www.widsconference.org/>) aims to inspire and educate data scientists worldwide, regardless of gender, and to support women in the field. Through Women in Data Science related events and many others we are able to connect to local tech community and engage allies and women in tech.



WOMEN IN DATA SCIENCE
DURHAM @GSK

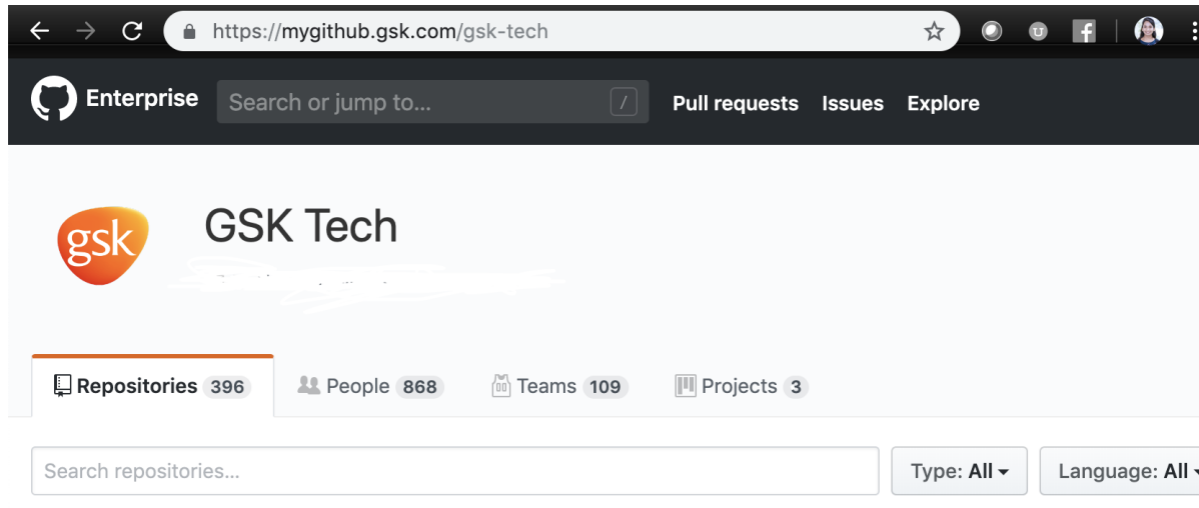
3. CODE REUSE THROUGH GITHUB AND COLLABORATION

One of our parameters to measure success is to look into the adoption of our provisioned tools and code re use. Our platforms to measure these are

- Code reuse through GitHub(<https://github.com/>)
- Collaboration through Workplace(<https://www.facebook.com/workplace>) Guilds

GITHUB

GitHub is a code hosting platform for version control and collaboration. It lets teams work together on projects from anywhere. In a very short span around 1000 programmers across GSK are currently using and sharing code in hundreds of repositories there by enabling code reuse across the enterprise



GUILDS

Guild(<https://medium.com/webcom-engineering-and-product/agile-guilds-the-yodle-way-47dc00f6cd3a>) is a group of people who work on different teams and meet with some frequency to discuss a specific competency popularized heavily by Spotify engineering culture(<https://labs.spotify.com/2014/03/27/spotify-engineering-culture-part-1/>) By leveraging internal social networking sites, we have active guilds which meet regularly to collaborate, learn from each other and share best practices and new ideas. Databricks guild is one such example at GSK. Databricks guild meets on monthly intervals covering topics collected through polls. We got great feedback from members and even management as it provided learning opportunity for those who are new to it and development opportunity to hone presentation skills for presenters. As it is volunteer driven, it hasn't been a burden on organizers as well.

CONCLUSION

Though we are in our initial stages of Digital Transformation journey, we are proud of the progress made so far. We are leveraging agile methodologies by collecting feedback, measuring progress and constantly refining our future strategy to increase speed and efficiency in democratizing Data science.

ACKNOWLEDGEMENT

I'd like to thank my amazing team. Most importantly Joshua Karstendick who was the technical lead; Core team members: Josh Davari, Elisa Orsini, Nick Peterson; Summer interns: Anna Darwish, Anjali Rajesh & entire GSK Data Product & Engineering leadership team and Platform engineering team (Special thanks to Gary Griggs, Mohammad Khalid & Kevin Harrigan).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Pragathi Kotha Venkata
 Company: GlaxoSmithKline
 City / Postcode: Durham/27709
 Web: <https://www.linkedin.com/in/pragathikv/>

Brand and product names are trademarks of their respective companies.